유입수량 예측을 통한 홍수기 댐 운영 효율화

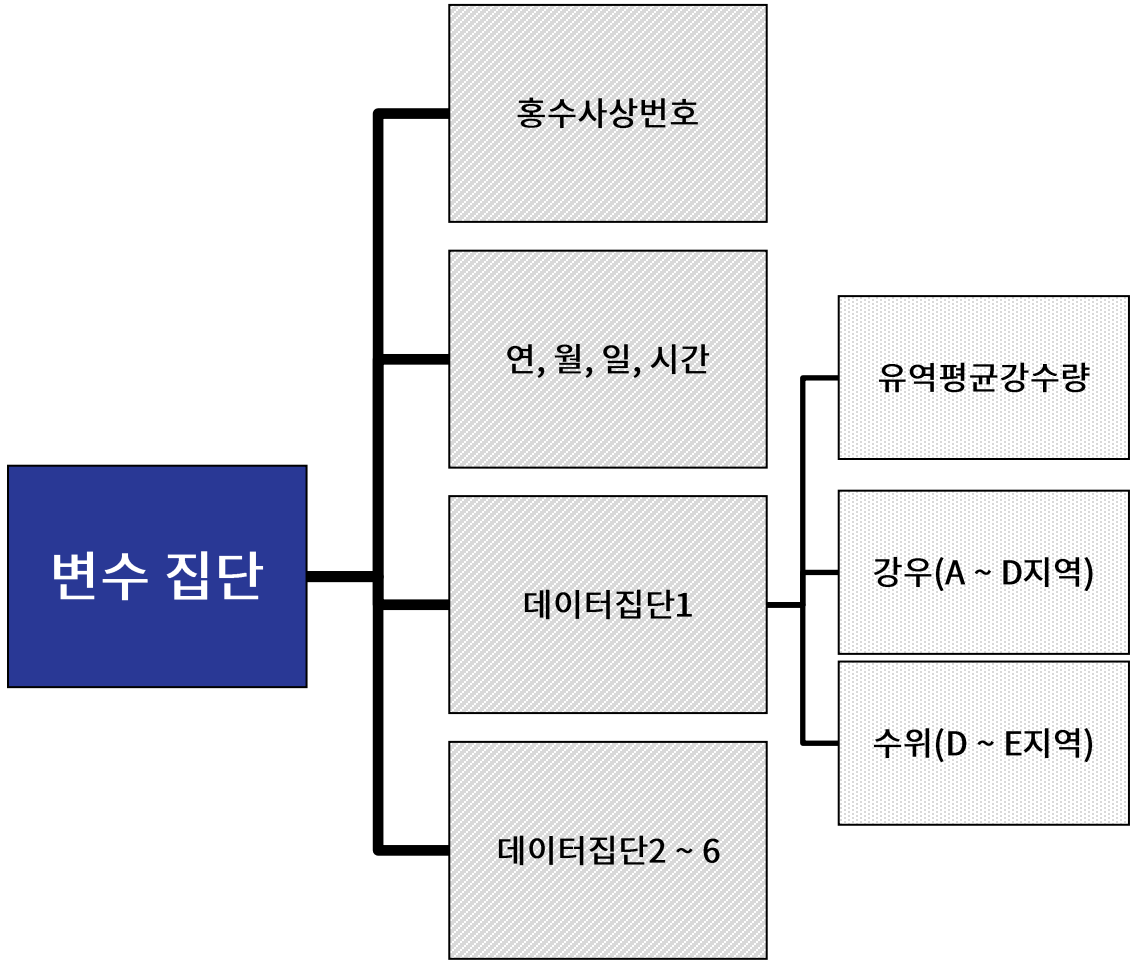
Team 공모전은 처음이라서
유동준, 박재연, 정은서

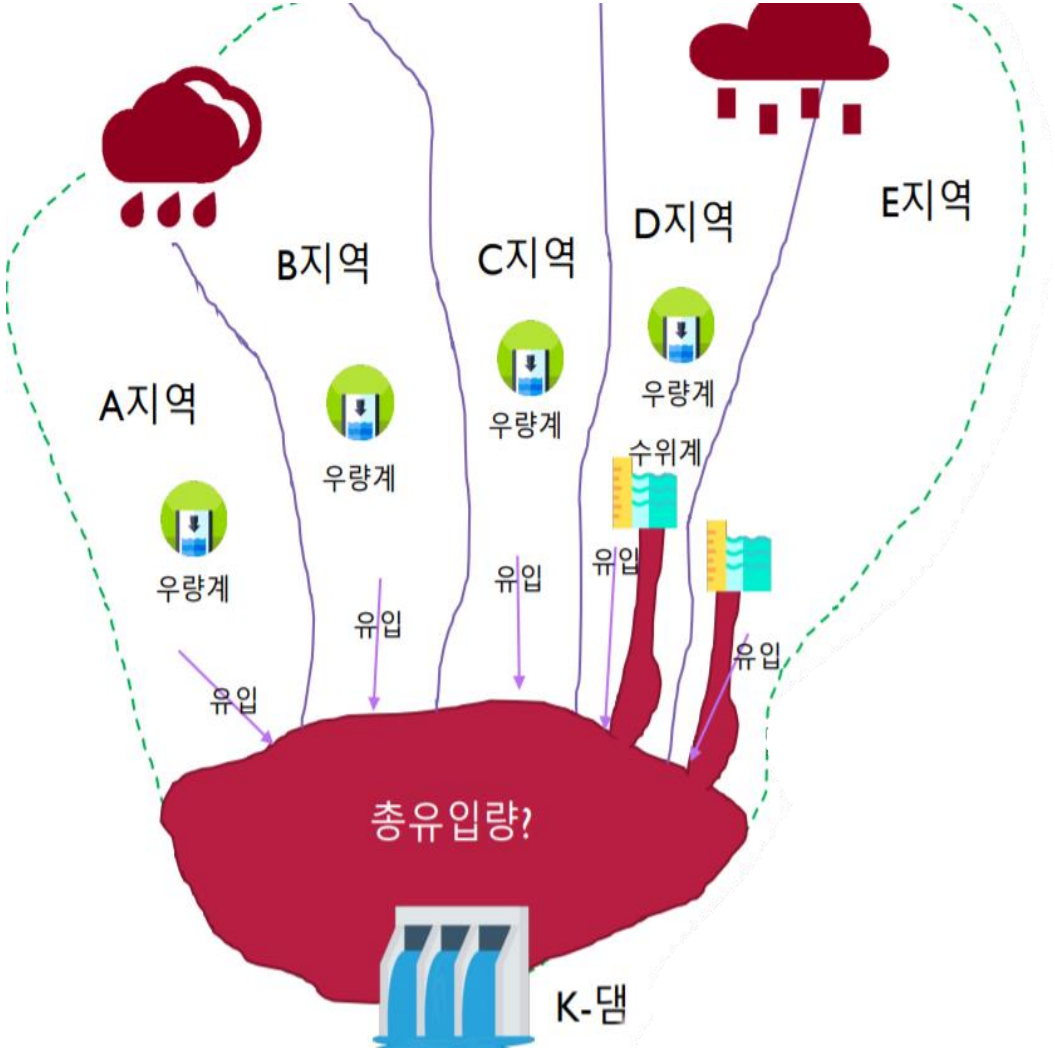
Contents

문제 정의

- 기존 홍수사상동안 데이터 → 특정 홍수사상의 유입량 예측
  ※ 홍수사상 : 강우가 상대적으로 커 토양/지반이 포화된 후 댐으로 많은 양의 유량이 흘러 들어온 기간(홍수가 발생한 기간)
- 저수량 = 유입량 - 방류량
- 방류량 결정에 도움 → 침수피해 예방

변수 탐색


변수 집단
홍수사상번호
연, 월, 일, 시간
데이터집단1
데이터집단2 ~ 6
유역평균강수량
강우(A ~ D지역)
수위(D ~ E지역)

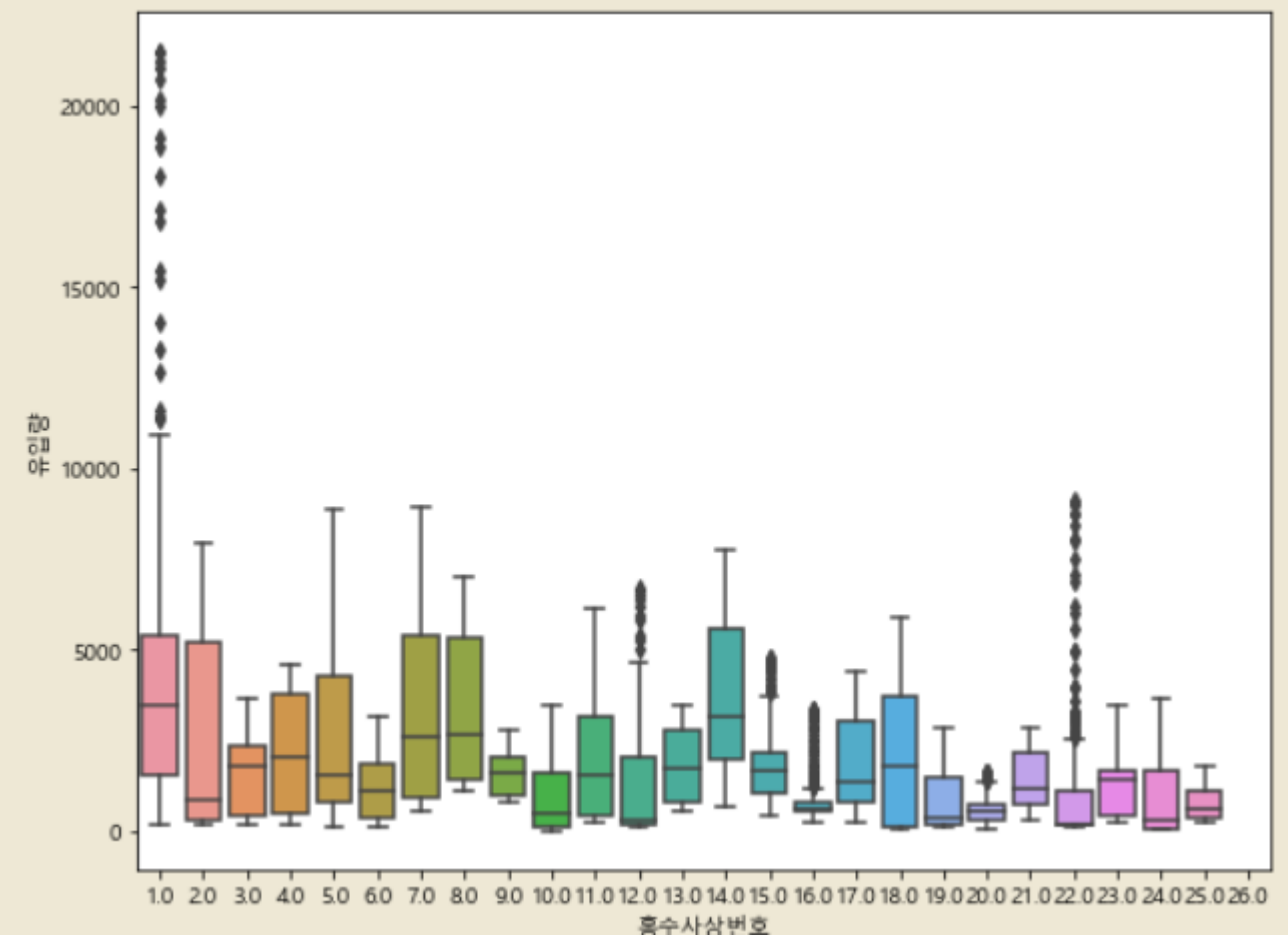

A지역
B지역
C지역
D지역
E지역
우량계
우량계
우량계
우량계
수위계
유입
유입
유입
유입
유입
총유입량?
K-댐

변수 전처리

- 유역집단 평균강수량과 A~D지역의 강우량으로 E지역의 강우량 추론
  - 주어진 유역이 한정된 점을 이용

- 각 변수를 유역집단 평균값으로 대체
  - ex) 집단1 – 강우(A지역)부터 집단 6 – 강우(A지역)까지를 강우(A지역)으로 대체
  - 총 43개의 변수에서 8개의 변수로 축소

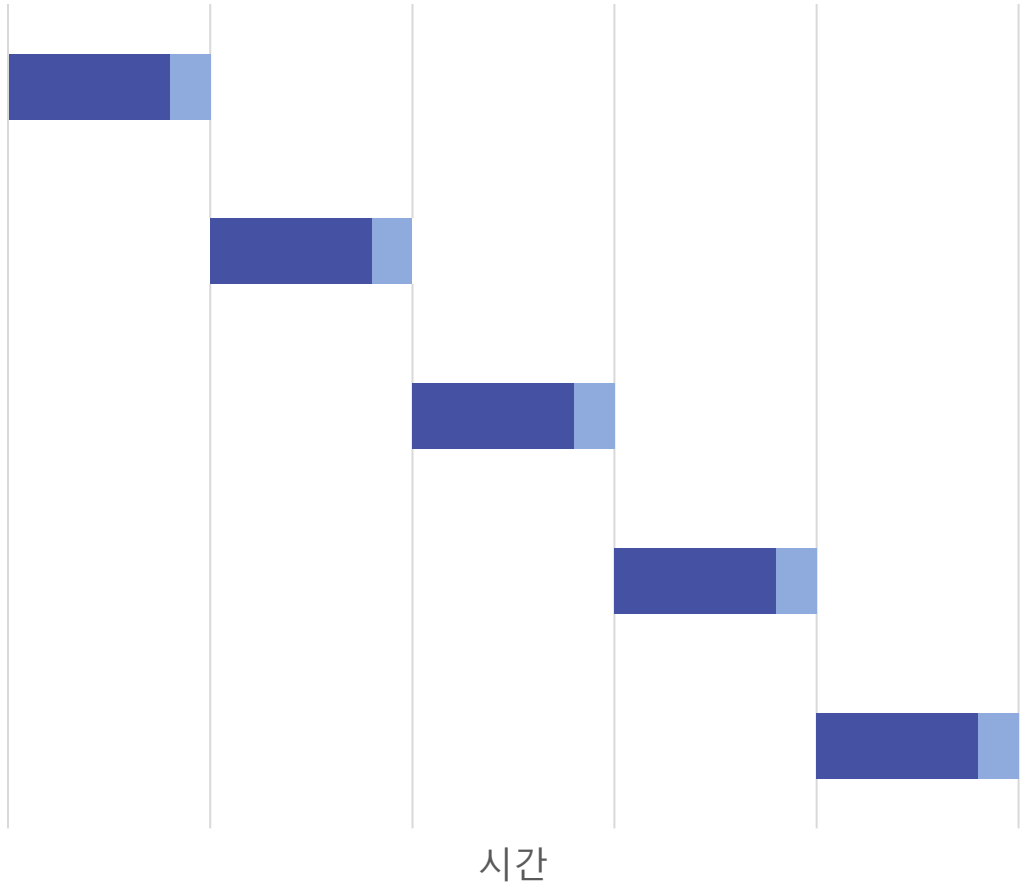
변수 전처리



이상치 분석

- 수위(E지역)이 이상치를 가지는 홍수사상은
  1, 12, 15, 16, 20, 22, 26
- 이상치의 원인 분석
  - 홍수의 분류 참고(수자원공사 제공)
- ‘태풍’이란 새 Feature로 활용
  - 총 9개의 변수와 1개의 유입량으로 정리

교차검증



장단점

- 장점
  - 시계열 데이터의 특성 보존
  - scikit-learn의 TimeSeriesSplit의 data leakage문제를 해결
- 단점
  - 연산량이 많음

교차검증

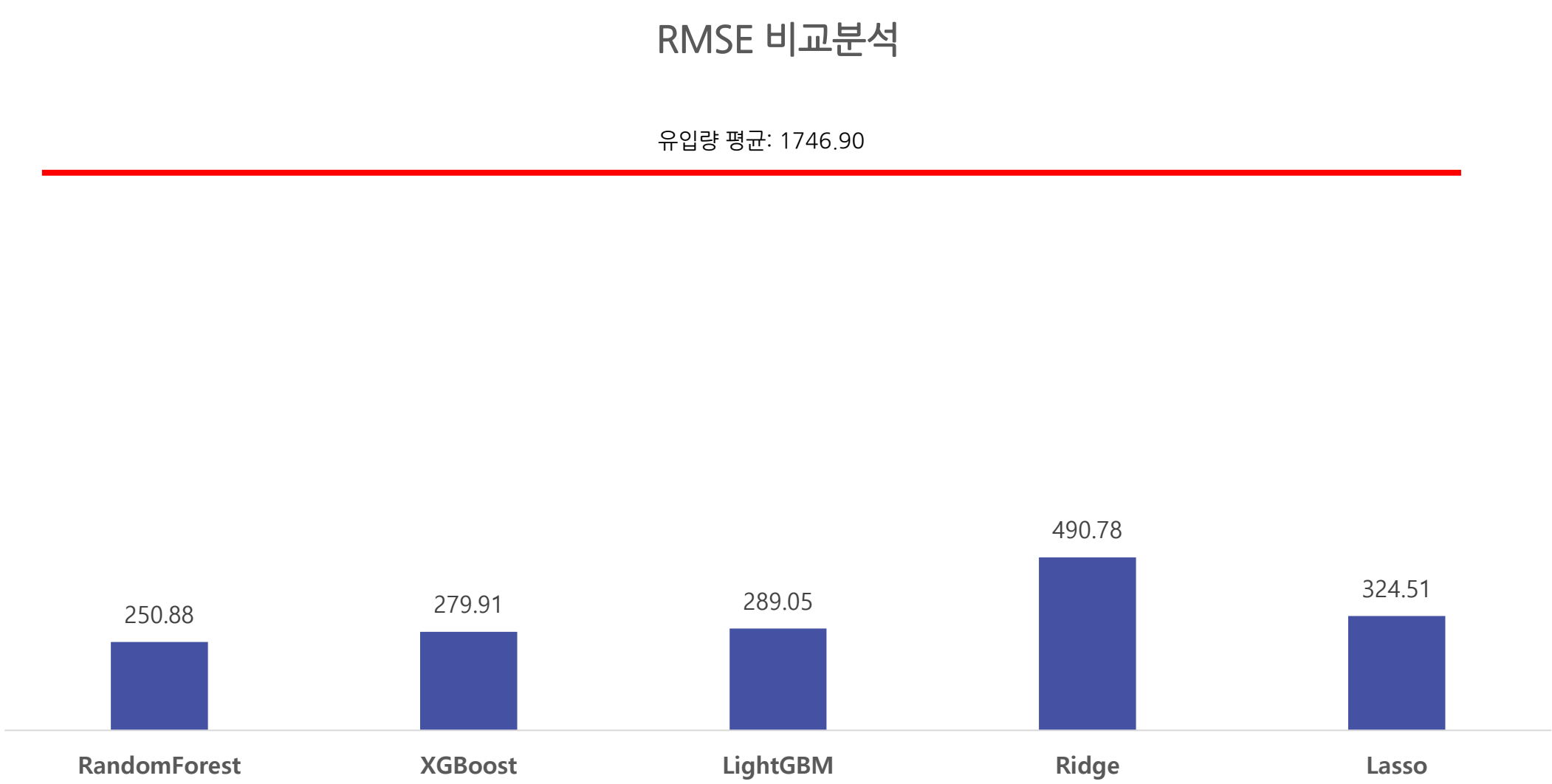
문제점

동일 길이의 구간으로 하면
이 데이터의 특성이 온전히 반영되지 않음

해결

FloodSeasonSplit() 객체를 만들어
홍수사상별로 fold별 훈련데이터를
나눠 교차검증 실시

모델링


RMSE 비교분석
유입량 평균: 1746.90
250.88
279.91
289.05
490.78
324.51
RandomForest
XGBoost
LightGBM
Ridge
Lasso

모델링

- VotingRegressor에 우수했던 모델 포함해서 앙상블 기법으로 결과 산출
  - RandomForest, XGBoost, LightGBM, Lasso

- 각 fold별 rmse결과

  rmse score : 1879.34
  rmse score : 556.16
  rmse score : 505.74
  rmse score : 261.34
  rmse score : 233.77

- 포함된 모델보다 우수한 결과를 보여줌

한계점과 보완(1)

한계

- 주어진 데이터셋만으로 분석을 진행해서 절대적인 양이 부족했다.

방안

- 장소가 특정된 데이터셋이었다면 외부 데이터를 연계해볼 수 있었을 것 같다.

한계점과 보완(2)

한계

- 딥러닝 모델을 학습시키지 못했다.
  - 데이터의 크기도 작고 시간이 연속되지 않기 때문

방안

- 정형데이터를 예측할 때 머신러닝 기법도 충분히 우수하며, 모델이 많을 수록 좋다고 생각하지 않는다.

Thanks to…

· “MyWater: K-water와 함께하는 물정보포털”, Retrieved October 18, 2021, from https://www.water.or.kr/realtime/sub01/sub01/dam/hydr.do?seq=1408&p_group_seq=1407&menu_mode=2

감사합니다